

Protein design: A perspective from simple tractable models

E.I.Shakhnovich

Harvard University, Department of Chemistry and Chemical Biology

12 Oxford Street, Cambridge MA 02138

(19 March 1998)

We review the recent progress in computational approaches to protein design which builds on advances in statistical-mechanical protein folding theory. In particular, we evaluate the degeneracy of the protein code (i.e. how many sequences fold into a given conformation) and outline a simple condition for "designability" in a protein model. From this point of view we discuss several popular protein models that were used for sequence design by several authors. We evaluate the strengths and weaknesses of popular approaches based on stochastic optimization in sequence space and discuss possible ways to improve them to bring them closer to experiment. We also discuss how sequence design affects folding and point out to some features of proteins that can be designed "in" or designed "out"

I. INTRODUCTION

The protein folding problem has two aspects: "direct" (i.e. folding) and "inverse" (i.e. "protein design"). The main issue of the "direct" PF problem is to understand the basic physical chemistry of how protein sequences determine their structure. The long-range goal of these studies is to predict protein conformation from sequence. The direct protein folding problem has received much attention recently and considerable progress was achieved, in understanding the general principles that govern folding of protein chains [1–4]. Using the language of bioinformatics one can define the folding problem as mapping the space of sequences into the space of structures.

The "inverse" protein folding problem is how to find a sequence that folds into and is stable in a given conformation at a given temperature. (see Fig.1). Again using the language of bioinformatics we can say that this corresponds to the mapping of space of structures to the space of sequences.

Fig.1

It is clear that the two problems are closely related to each other: better understanding of the principles of protein folding makes it possible to clarify which features of protein

sequences are necessary (as well as sufficient) for their stability and fast folding, i.e. what makes protein a protein. Such understanding focuses the attention of designers on emphasizing those crucial features of folding sequences.

The experimental approaches to protein structure determination have been very successful providing a wealth of structural information. While the growing flow of genomic information makes the development of theoretical approaches to predict protein conformation even more desirable, there is an experimental "shortcut" of X-ray crystallography or NMR to the solution of the "direct" PF problem.

The situation with design is very different. Most of the present experimental approaches enjoyed only limited success providing polypeptides which in most cases fold into compact but mostly disordered conformation of molten-globule like species (see e.g. [5]). It is quite possible that limitations in experimental design are due to relatively low synergism between experiment and theory in that area. An important success story based on such synergism of theory and experiment is given in [6] where theoretical analysis has helped to guide the design effort which resulted in a small protein that folded into predicted "target" conformation. This work clearly demonstrates a crucial role of theory in protein design. A limitation of the approach reported in [6] is that it requires complete enumeration of sequence candidates - a problem that explodes exponentially with chain length and thus limits this valuable approach to relatively short lengths. The success and limitation of the work of Mayo and coworkers call for further refinement of theoretical approaches to protein design some of which will be outlined in this review.

It is important to note that the bottleneck in protein design is not on the synthetic side, but rather in the fundamental problem that researchers generally do not know which sequences to synthesize. Since the number of possible sequences is enormous, and the fraction of them that are able to fold into protein-like structures is negligible (see below) the probability to "hit" a correct sequence by chance is vanishingly low. Of course there exist clever experimental approaches, like phage display [7] which bias experimental sequence search towards better candidates. However, in our view, convincing success in protein design will

come with reliable theoretical approaches which will make it possible to find sequences that fold uniquely into a desired conformation. Perhaps this goal alone justifies all the effort that has been put into protein folding theory over last few years.

In this review I will discuss how recent advances in understanding protein folding help us in the efforts to design protein sequences and understand their natural evolution.

II. MAPPING STRUCTURES INTO SEQUENCES: HOW MANY PROTEIN SEQUENCES ARE THERE?

The computational approach to protein design aims to find sequences that fold to a given structure, in a particular model. The fundamental question, is if there is any solution to this problem (for a model of course, we know that there is one for proteins) and if yes, how many solutions are there, i.e. how many sequences can fold into a given conformation. This question can be addressed only if we understand what features should a folding sequence have. Such understanding builds on recent developments in protein folding theory which elucidated some of the properties of folding sequences [8–11].

According to thermodynamic hypothesis [12] sequences that fold into a given structure have lowest energy (potential of mean force) in that structure, compared to energies of decoys, i.e other conformations for that same sequence. The "consistency principle" due to Go [13] and "principle of minimal frustrations" (PMF) by Bryngelson and Wolynes apparently posited that the necessary condition for protein stability and fast folding is that the native state has energy that is much lower than energies of the bulk of misfolded states (decoys). Speaking modern language one can say that PMF is actually equivalent to the requirement of large energy gap in protein-like models

The results of analytical microscopic theory of heteropolymer folding [14–17] as well as numerical studies [10,9,18] in lattice model are consistent with the PMF. More specifically, it was shown that in order for a sequence to fold into a given native structure, its energy in that structure should fall below a certain threshold E_c . E_c is the energy at which the

density of states for decoys vanishes: at $E \geq E_c$ the density of states is very high so that many decoys belong to that energy range (see insert in Fig.2). The probability that there will be a decoy, structurally unrelated to the native conformation and having energy $E < E_c$, has been estimated in the Appendix to [10] to be $\exp((E - E_c)/T_c)$, where T_c is the temperature of thermodynamic freezing transition in random heteropolymer. (The thermodynamic freezing transition is defined as temperature at which entropy of a polymer vanishes [14,19]). Therefore if a sequence folds into a given structure with energy E , the probability that there will be structurally dissimilar decoy having equal or lower energy falls off exponentially and for sequences that fold into the target structure with sufficiently low energy E such that $E - E_c \gg T_c$, the target structure will almost certainly be a unique ground state conformation. Fig.2

Further studies showed that pronounced "stability gap" $E - E_c$ is also sufficient to provide fast folding for lattice model proteins of considerable length (more than 100 monomers) [18,20], consistent with the PMF [21].

Therefore a possible search criterion for folding sequences is large (many kT_c) stability gap. With that the issue of how many sequences can fold into a given conformation (degeneracy of the protein code) is reduced to the question of how many sequences $\mathcal{N}(\mathcal{E})$ exist that have energy $E < E_c$ in a given structure:

$$\mathcal{N}(\mathcal{E}) = \sum_{f \in \Pi} \delta(\mathcal{H}(f \mid \Pi, \Pi \setminus \{f\}) - \mathcal{E}) \quad (1)$$

Where $H(seq, conf)$ is energy of a particular sequence in the target conformation. Delta means that summation is taken over all sequences that have energy E in the native conformation. A particular example which got much attention in the past [22–24,1,25] is when H is a contact potential:

$$H(\{\sigma\}, \{r\}) = \sum_{i < j} (U(\sigma_i, \sigma_j)) \Delta(r_i, r_j) \quad (2)$$

where N is the number of residues in the chain. The symbol σ_i characterizes the type of monomer i so that sequence of monomers is defined as sequence of symbols $\{\sigma\}$. There

are 20 types of aminoacids so that $\sigma_i = 1...20$. The parameters $U(\sigma_i, \sigma_j)$ determine the magnitude of contact interaction between monomers of type σ_i and σ_j ; several sets of such parameters were published ([22,23,26,27]). A simple approximation of conformation of a chain is residue representation whereby a residue i is assigned a one point location variable r_i (it can be a geometrical center of the side-chain or coordinate of its C_α or C_β atoms).

$\Delta(r_i, r_j) = 1$ if residues i and j are in contact and 0 otherwise. For protein structures a reasonable definition of a contact is when distance between their C_α/C_β atoms is less than 6.5Å ([22]). For lattice model proteins definition of a contact is even simpler: two aminoacids that are lattice, but not sequence neighbors are considered contacting.

$\mathcal{N}(\mathcal{E})$ in eq.(1) can be evaluated using the technique that represents Dirac delta-function in eq.(1 via Fourier transform, expands appearing exponentials up to the second order, sums over all sequences and re-exponentiates the result. The final result of the calculation can be expressed in terms of "entropy" in the sequence space:

$$S_{seq}(e) = \ln \mathcal{N}(\mathcal{E}) = \log(\mathbb{1}_{\{\mathcal{E}\}}) - \frac{(\mathcal{E} - \mathcal{E}_{-})^\epsilon}{\epsilon \mathcal{N} \mathcal{D}^\epsilon} \quad (3)$$

m_{eff} is the effective number of types of aminoacids:

$$m_{eff} = \exp\left(-\sum_{i=1}^{20} p_i \ln p_i\right) \quad (4)$$

(e.g. if all types of aminoacids are equally represented so that $p_i = 1/20$ for any i then $m_{eff} = 20$. In the opposite case when, say $p_1 = 1$ and $p_i = 0$ for any $i = 2...20$ then $m_{eff} = 1$ which makes clear sense since the latter situation corresponds to a homopolymer.) E_{av} is average (over all conformations) energy of interactions, per aminoacid. and D is the dispersion of interaction energies (per contact). E_{av} is calculated as an average interaction energy over all *possible* contacts; It depends on aminoacid composition but not on details of the sequence. D is dispersion of contact energies also calculated over all *possible* contacts. Calculation of these quantities does not require simulations or enumerations in conformational space. However, certain geometrical properties which may restrict the types of possible contacts should be taken into account. can be taken into account. For example, for a cubic lattice

an important property is that only possible contacts are between units with odd and even positions along the chain. This "even-odd" rule should be taken into account in estimate of E_{av} and D for cubic lattice model.

(The question of how many sequences fold into a given structure was first addressed by Finkelstein, Badretdinov and Gutin who postulated the distribution given in eq.(3) [28]).

According to the heteropolymer theory [21,14,29,19] the density of states of 3-dimensional heteropolymer (the number of conformations having energy in a given range) follows the Random Energy Model distribution:

$$W(E) = \gamma^N \exp\left(-\frac{(E - E_{av})^2}{2ND^2}\right) \quad (5)$$

The energy at which the chain runs out of states (the boundary of the continuous spectrum E_c in the insert in Fig.2) is estimated from the condition $W(E) \sim 1$, i.e.

$$E_c - E_{av} = N(2 \ln \gamma)^{1/2} D \quad (6)$$

As explained above, a necessary condition that determines a folding sequence is that its energy in the native state is $E < E_c$. Such sequences should exist, i.e. $S_{seq}(E < E_c) > 0$. It follows from 6 and 3 that this condition can be satisfied only when

$$m_{eff} > \gamma \quad (7)$$

Apparently, there is another threshold energy, E_{lowest} such that there are no sequences that have energy in the native state lower than E_{lowest} . A possible crude estimate of E_{lowest} can be obtained from the condition that at this energy the system runs out of *sequences*. Mathematically this is equivalent to the condition $S_{seq}(E_{lowest}) = 0$. However it is quite possible that this is an overestimate and the actual boundary of lowest possible energies in a sequence model may be higher than estimated from the entropy condition below.

Therefore, the upper bound estimate of the maximal possible gap $E_{lowest} - E_c$ is

$$G_{max} = N \ln \frac{m_{eff}}{\gamma} (2D^2)^{1/2} \quad (8)$$

Fig.3

A specific simple example to clarify the main concepts of this analysis is presented in Fig.3. It shows the energy spectra, or densities of states (log of the number of conformations having a given energy) for the designed (black bars) and a random sequence having the same composition (13B, 14W) (grey bars). Comparing this spectrum with the one presented schematically in the insert in Fig.2 one should keep in mind that for the model that has only two kinds of aminoacids the spectrum is apparently discrete because possible values of energy are determined by numbers of contacts of different kinds which are obviously integer (a straightforward generalization of heteropolymer results to this discrete case is given in [30]). However, the occupancy at each energy level (i.e. how many conformations have that energy) is different for different levels. Specifically, there may be energy levels that are highly populated i.e. a multitude of conformations have that energy. There exist also empty low-energy levels which can be filled only for special sequences (i.e. only special sequences can have such an "unusually" low energy in their native conformations). The designed sequence shown in Fig.3 has absolute lowest possible, for the model, energy $E_N = E_{lowest} = -84$ in its unique native conformation.

It can be seen clearly in Fig.3 that the spectra for the random and the designed sequences differ only at the low energy part: at energies that are higher or equal than -60 both random sequence and the designed one have almost identical spectra, i.e. this part of the spectrum is sequence independent (quantities that are sequence independent are called self-averaging [31,29,2]). According to the heteropolymer theory [14,29,21,19] the density of states is self-averaging at energies E_c and higher while the low-energy part at $E < E_c$ is sequence specific. The low-energy non-self-averaging part of the spectrum represents an energetic fingerprint of a sequence.

It follows that for this model $E_c = -60$. Note also the concave shape at the left wing of the spectrum for designed sequence which is a signature of a cooperative transition [13]. The cooperativity of transition (e.g. its widths) is directly related to the value of the relative gap $g = (E_N - E_c)/E_N$. For this model $m_{eff} \approx 2$. Only compact conformations are considered, therefore $\gamma = 103346^{1/26} \approx 1.7$. The relative gap is $g = -0.33$.

III. LESSONS FOR DESIGN

The statistical-mechanical analysis suggests a number of lessons.

Lesson1: The design problem may be easier than folding problem. In a protein-like model where $m_{eff} > \gamma$ there is an exponential in chain length N number of sequences that have sufficiently large energy gap $G \sim ND$ to fold reliably into the target structure. Unlike folding where a unique ground state solution is sought, in design any sequence having sufficient (not necessarily the greatest possible) energy gap [8,9] folds cooperatively into the target conformation if the temperature is not too low, see [32]. While the number of folding sequences is large, the *fraction* of folding sequences (i.e. the probability to pick up a cooperatively folding sequence from the ensemble of random sequences) is quite low. That makes the design problem nontrivial.

Lesson2: "The number of types of aminoacids may be an important factor that determines the designability of a protein model"

The models where the number of types of aminoacids m_{eff} is small are "undesignable". This means that even the best sequences designed for these models have energy in the native state higher than E_c , i.e. decoys with energy lower or equal to the energy of the designed sequences in the native state are present in such models. Apparently no folding is possible in this case since the native structure is not unique. An example of such undesignable model is the so-called HP model [33].

Lesson 3: "Stiffer" chains provide greater energy gaps and therefore are more designable

The fundamental relation for a designable model, the condition presented in (7) can be enforced either by increasing the number of aminoacid types or by decreasing γ i.e. by decreasing the number of conformations (per monomer). There is a number of ways to decrease γ : formation of secondary structure, forcing the conformational ensemble of a chain to the set of compact conformations (by introducing additional non-specific attraction, Fig.3), biasing the conformations to carry certain structural features (like in threading). The example given in Fig.3 shows that even the "two-letter" model may sometimes have non-

degenerate native state (but very small gap) if its configurational space is restricted to only compact conformations. When full ensemble is considered the ground state of HP sequences become multiple degenerate [34,9,33]. Apparently the number of all conformations (per monomer) γ_{all} is greater than the number of compact conformations $\gamma_{compact}$ so that the condition (7) is violated for the HP model when all conformations are considered. On the other hand the "two-letter" models that are restricted to maximally compact conformations only are just "on the borderline" of the validity of the condition (7).

Lesson 4: Protein design for most 3-dimensional models does not require "designing out" the decoys; 2-dimensional models behave very differently and require more complicated design that may require "designing out" the decoys.

The key to successful protein design is to find sequences that have low energy of the native state without optimizing decoys at the same time. This factor increases the energy gap or, equivalently, increases the thermal probability to be in the native state (see below). To this end the "ruggedness" of the conformational space of 3-dimensional random heteropolymer (as exemplified by the equivalence between heteropolymers and the Random Energy Model (REM) [21,14,15]) plays a key role. According to the REM, most low-energy decoys are structurally different from the native state (except the ones that represent small fluctuations around the native conformation - the native state ensemble). To this end optimization of the native conformation energy (i.e. making the native contacts stronger) does not affect the low-energy structurally dissimilar decoys (see Fig.3). That makes the designing "in" on the background of decoys that are unaffected by sequence selection efficient to increase the gap. We should emphasize that this is true only for 3-dimensional models; in two dimensions the optimization of the native states gives rise to optimization of numerous partly folded low-energy decoys making the native state unstable (in contrast to the 3D case where partly folded decoys have high energy). The physical reason for such dramatic dependence on space dimensionality, is given in [35,36] (especially see appendix to [36]): In 3-dimensional compact chains non-local contacts dominate while in 2-dimensional chains local contacts play are dominant.

It was pointed out by several authors [37–39] that some special 3-dimensional target conformations (crumpled globules [40]) may be as "undesignable" by simple methods as two-dimensional models, for the same reason - prevalence of local contacts.

IV. STOCHASTIC OPTIMIZATION IN SEQUENCE SPACE: SIMPLE MODEL SOLUTION FOR THE DESIGN PROBLEM.

The major lesson from the statistical mechanical theory is that many solutions of the design problem exist. A crucial question of practical importance is how to find such solutions. To this end a number of approaches, (reviewed in this chapter) of various complexity and efficiency have been suggested.

It is clear that all what is needed for successful design is to find a sequence $\{\sigma_i\}$ that has high thermal probability to be in the native state:

$$P(T) = \frac{e^{-\frac{H(\{\sigma_i\}, \{r_i^0\})}{k_b T}}}{Z(\{\sigma_i\})} \quad (9)$$

Where the native state is characterized by the set of coordinates of its residues $\{r_i^0\}$, H is the energy of a given sequence in a given conformation (cf(2)). Z is a partition function of the chain

$$Z(\{\sigma_i\}) = \sum_{r_i} e^{-\frac{H(\{\sigma_i\}, \{r_i\})}{k_b T}} \quad (10)$$

where summation is taken over all conformations of the chain $\{r_i\}$. T is temperature and k_b is Boltzmann constant.

As presented by eqs.(9,10) the problem of design is of great complexity since it involves search in both conformational and sequence spaces. (The search in conformational space is needed to determine the partition function). In other words the "exact" solution of the design problem that includes exhaustive searches in conformational and sequence spaces would require $(m_{eff}\gamma)^N$ "trials" - a prohibitive number for any model of practical interest.

This calls for development of approximations that would allow to avoid exhaustive search both in sequence space and in conformational space. A simplest approach of this kind was proposed in 1993 in [9]. It is based on the following ideas:

i) The optimization of stability is equivalent, in a simplest case, to the maximization of the energy gap g defined above (see Fig.1 of [9] for a qualitative explanation of this fact). The boundary of the continuous spectrum E_c is a self-averaging quantity, i.e. it depends on aminoacid composition only while the lower part of the spectrum $E < E_c$ is highly sequence specific. This conjecture from heteropolymer statistical mechanics was shown to be correct for simple exact models, such as the one shown in Fig.3. It follows that the desired design results can be obtained by selection of sequences that have low energy in the target conformation at a given aminoacid composition. It is clear that this statement is equivalent to the assumption that the partition function Z (more precisely contribution to Z from non-native-like decoys) in the eq.(9) depends primarily on aminoacid composition rather than on sequence. The analysis using the Random Energy Model approximation suggests that this conjecture is valid at high enough temperature $T > T_c$ where T_c is temperature of the "freezing" [21,14,19] transition in a random heteropolymer having the same aminoacid composition. A lucid discussion of this point and further details can be found in [19].

The gap optimization in sequence space can be achieved by any stochastic algorithm. In the case of sequence design the energy landscape in sequence space is "smooth" [9,41], so that there is no complicated search problem. Therefore a simple Monte-Carlo algorithm would suffice [9,18,42,20].

An *experimentum crucis* to test the statistical-mechanical approach to sequence design is to pick an arbitrary conformation and design a sequence that is expected to fold into that conformation. A proof of concept for a design method is an actual folding simulation of a designed sequence, starting from an arbitrary random coil conformation. If the designed sequence converges to the target conformation and never encounters grossly misfolded conformations with energy lower than the target conformation then they may be stable in the target state, and the design is successful.

This program has been carried out in [9,18] where random mutations preserving the aminoacid composition (monomer swaps) were introduced under Metropolis control with certain "selective" temperature T_{sel} . The model studied in [9] is the same as shown in Fig.3. Strong attraction between any pair of aminoacids shifted the conformational ensemble in folding simulations towards compact states. The designed sequences were shown to fold into the target (native) conformation which in all cases turned out to be the non-degenerate global energy minimum.

An attempt to carry out a rigorous test of design for longer sequences (48-mers) in the HP model without introducing strong overall attraction was not successful: In that case the native conformation was always multiple degenerate. The non-compact decoys often had lower energy than the target conformation. These results are consistent with earlier prediction [18] and the presented statistical-mechanical analysis.

Therefore the two-aminoacid type model design cannot be successfully extended to longer chains because of the requirement to restrict the conformational ensemble by compact conformations only (see Fig.3). Introduction of non-specific additional attraction to bias the conformational ensemble towards compact conformations dramatically slows down folding making it infeasible to fold longer chains [43,32,44]. Thus the range of lengths that can be studied using the two-aminoacid type model is very limited. Such limitation may give rise to some small-size artifacts.

An obvious solution of this problem is to use a greater number of kinds of aminoacids than just two. This was done in [18] where 20 types of aminoacids and Miyazawa-Jernigan interaction potentials [22] were used. The design-folding program was carried out for 20-aminoacid type model proteins on a cubic lattice (with fixed composition corresponding to an "average" aminoacid composition in proteins). The designed sequences of 80-mers folded fast and were stable in their target conformation; No conformations with energy lower than the energy of the target conformation (for the designed sequence) were encountered. These results provided, for the studied model, an important proof that design approach based on statistical-mechanical theory of protein folding is feasible and is basically correct, for the

right model.

A somewhat different interesting approach to design was proposed by Grosberg and coworkers [42,20]. This approach is based on the idea of pre-biological evolution by "imprinting", according to which first macromolecules could have evolved as a result of polymerisation of equilibrated monomers which could have interacted with substrates at pre-polymerisation stage. The "imprinting" design procedure also uses the MC annealing protocol but in the system of disconnected aminoacids. After that the chain is threaded through the "annealed" configuration of monomers on the lattice, thus creating a sequence. The advantage of this method compared to the design procedure proposed earlier in [9,41] is that it can be (in principle) experimentally realized in an abiotic system. A disadvantage is that sequences obtained by "imprinting" are considerably less stable in their native conformation and sometimes they may even not have the target conformation as global energy minimum. The reason is that sequence design uses the energy function in which nearest neighbors in sequence do not interact (their interaction adds a constant to energy of each conformation and therefore it is irrelevant). The imprinting method does not take this factor into account, therefore when a chain is threaded through the annealed system of monomers it will often connect strongly interacting nearest neighbors, making them covalently bound and therefore losing their strong attraction for stability of the native state. Despite of that difficulty it was demonstrated that the sequences obtained as a result of imprinting procedure are often able to fold into their native conformation corresponding to global energy minimum [42,20].

Several authors proposed other, than MC optimization techniques to search sequence space [45,46]. In our opinion, the MC search in sequence space is as efficient as other optimization algorithms (because the landscape is smooth and multitude of solutions exist). However, the MC approach is advantageous because it converges to the canonical distribution and hence its results can be rationalized from the statistical mechanical perspective.

This interesting analogy between the statistics in sequence space and several statistical-mechanical models was noted in [9,41,18,47]. The Hamiltonian for sequence design eq.(2) (where the coordinates are quenched but the aminoacid identity variables σ are allowed

to vary) is analogous to the Hamiltonian of the Ising model if there are only two types of aminoacids and to the Potts model if there are many types of aminoacids. It was pointed out in [9,41] that the MC design procedure converges to the canonical distribution in sequence space. Therefore the statistics of sequences become analogous to the statistics of "spin configurations" in the equivalent statistical-mechanical models as it follows the same Boltzmann law. This analogy is explained in more detail in [41] where the one-to-one correspondence between statistical characteristics of sequence design and Ising model are listed in the Table 1. (Two-aminoacid type sequences were considered in [41] but the results are trivially generalizable to the multi-aminoacid type models).

Of those analogies probably the most important one is the relation between entropy in statistical-mechanical models, and "degeneracy" of the protein code. This analogy allows us to calculate $\mathcal{N}(\mathcal{E})$ directly from the MC sequence design simulations. The idea of the calculation is based on the thermodynamic equation that relates the entropy at a given temperature T with average energy at the same temperature via:

$$S(T) - S(\infty) = \frac{E(T)}{T} - \int_T^\infty \frac{E(t)}{t^2} dt \quad (11)$$

with $S(\infty)$ being entropy of a system at infinite temperature. In our case of sequence design, the selective temperature, at which MC design procedure in sequence space is carried out, plays the role of temperature in eq.(11). $S(\infty)$ corresponds to random sequences without a bias towards any particular structure. $S(\infty) = N \ln m_{eff}$. The results of the calculation are shown in Fig.1 for several proteins with the energy function approximation given by eq.(2) (the sequence design simulations for each protein in Fig.1 were carried out keeping the aminoacid composition fixed and equal to the aminoacid composition of native sequence for each protein see [9,41])). (The related results were presented in a recent publication [47]). The solid line in Fig.1 shows a theoretical estimate given by the eq.(3). It is quite clear that the theoretical estimate is in excellent agreement with the simulation results. Further, it is clear from Fig.2 that sequence entropy, is approximately the same for all studied proteins (of course different sequences fold into different protein structures; it is the *number*

of sequences that is invariant for different proteins). Such invariance is understandable since in this approximation the difference in energy functions eq.(2) between proteins are due the average coordination number of their aminoacids and the connectivity, i.e which of the spatially proximal aminoacids are sequence neighbors. While these factors are crucial in determining which sequences actually fold into a given conformation, they are not too specific to give rise to pronounced differences in "designability". This result of the analysis of the model with 20 types of aminoacids can be compared with the "designability principle" suggested by Finkelstein and co-authors [48] and further addressed by Tang and co-authors [49]. The analysis presented in Fig.2 differs from that of Finkelstein et al that we did not impose energetic penalties on certain structural features such as turns etc while these factors were assumed to be important in [48]. On the other hand the arguments presented in [48] are the phenomenological ones that assume a certain form of density of states for a particular structure; the justification of such assumptions based on a more microscopic model will be very interesting to obtain.

Tang and co-authors used a standard 27-mer models [50] with the form of energy function similar to eq.(2). These authors carried out exhaustive enumeration of all compact conformations and all "two-letter" sequences. The "designability" of a structure was defined in [49] as the number of sequences that have this structure as a unique energy minimum among all compact conformations. Interestingly Tang et al report that certain structures of compact 27-mers are more "designable" than others in their model. Further they infer that the designable structures feature protein-like properties such as secondary structure.

It follows from the present analysis that the issue of "designability" may be indeed important for the models that feature two kinds of aminoacids because some structures can accommodate their "best" (lowest energy) sequences with slightly lower energies than other structures can accommodate their "best" sequences. In the situation when there is no significant gap, this small energy difference between different structures matters a lot: a more designable structure can accommodate their sequences with energy slightly lower than E_c while less designable ones may have E_{lowest} that is close or above E_c . These factors can be

clearly seen in Fig.3. For the structure shown there the sequences with lowest possible energy $E_{lowest} = -84$ exist. The lower the energy of the native state is the lower the probability that a decoy having the same energy will be found (see above and [10,30]). Correspondingly there may be many sequences that have the structure shown in Fig.3 as their *unique* ground state, i.e. this structure may be highly designable. It is clear that the designability of this structure is due to the special pattern of bonds on the lattice which makes it possible to find a sequence that features complete separation between beads of opposite kind (sequence neighbors do not interact). However, there are many structures that do not have such an "ideal" pattern of bonds so that even their "best" sequences still have at least one contact between aminoacids of opposite kind. For them $E_{lowest} = -82$. For those sequences the gap is smaller and therefore they are less designable than the structure shown in Fig.3. This is consistent with the observation of Tang and coworkers that more designable structures deliver greater energy gaps [49].

This analysis implies that the pronounced difference in designability exist for the models where even the maximal possible gaps are small (i.e $m_{eff} \approx \gamma$). In that case every favorable contact matters a lot so that differences between structures (patterns of bonds on the lattice) which allow to gain or lose an extra favorable contact may make a significant impact on designability. In many aminoacid kinds 3-dimensional models where sequences can have energy in a target conformation that is considerably below E_c (i.e. $m_{eff} > \gamma$) all structures may be highly designable. Therefore it is important to extend the study of [49] to multi-aminoacid type model. However, such extension is a difficult one: It is computationally very costly to enumerate the multi-letter sequences exhaustively as it was done for two-letter sequences by Tang and coworkers [49]. The MC simulations in sequence space may be a reasonable alternative to exhaustive enumeration of sequences. The results presented in Fig.2 show no visible differences in designability for a few protein structures which were used for the analysis.

An important caveat of the MC sequence analysis should be mentioned here. The estimate of the number of sequences in eq.(11) is based on the thermodynamic analogy which

is not precise enough to take into account sub-dominant (in N) contribution to entropy in sequence space. Therefore, though the major, exponential in chain length, contribution to the number of sequences that fold into a given structure (corresponding to the linear in N contribution to sequence entropy), is the same for different proteins, there may be sub-dominant (less than exponential in chain length) contributions which may give rise to some differences in designability. Whether this is so and if yes, whether this is important for our understanding of protein evolution is a matter of future research.

The approach to the design which uses MC simulation in sequence space with fixed aminoacid composition [9,41,20] is simple, computationally very efficient and is non-heuristic one (i.e. it is not limited to any particular model of a protein). Hence its appeal.

However, it has certain disadvantages most important of which are:

a) Keeping the aminoacid composition fixed eliminates the possibility to find an optimal (for folding and stability) aminoacid composition.

b) The assumption of sequence independence of the partition function in eq.(9) (more precisely the contribution to it from non-native decoys) follows from the mean-field heteropolymer theory [14,19]. However, this assumption is valid only at high temperature. Furthermore, the deviations from the mean-field predictions need to be examined.

c) The lack of reference to the temperature at which sequence is expected to fold. Indeed, in the full design problem sequence space optimization of $P(T)$ in the eq.(9) both the numerator and denominator depend on temperature and it is possible that at different temperatures different factors become important to optimize.

Those limitations were partially overcome in a number of subsequent publications [36,51–53].

The first limitation (constant aminoacid composition) was overcome in [36,54] where the quantity $Z = (E_N - E_{av})/D$ (the so-called Z -score, [55]) was optimized in sequence space.

Optimization of the Z -score instead of native energy fixed one of problems of the simple approach [9,41] - convergence to homopolymeric sequences unless the aminoacid composition is constrained. As a result, the design based on optimization of the Z -score was able to find

also optimal composition which provided the best value of the gap.

A number of recent papers [51–53] addressed the second problem, attempting to better estimate the partition function Z than simply assuming it to be sequence-independent. In general this problem is very complicated since an exact solution would require enumeration of conformations after each mutation (to evaluate Z for the new sequence) which makes it computationally very difficult for small chains and totally prohibitive for longer chains of realistic length.

The paper [53] attempted to optimize directly $P(T)$ in eq.(9) using dual Monte-Carlo: in sequence and conformational space (chain growth algorithm was applied for conformational space simulation). This approach requires considerable computational effort in order to reach Boltzmann distribution to provide a correct estimate of the partition function Z . Even for shorter chains such equilibration would require more than 10^5 MC steps and this number grows fast with chain length [56] making the interesting approach proposed by Seno et al [53] very demanding computationally. The apparent advantage of this approach is that it contains direct reference to folding temperature and is rigorous. The disadvantage is that it is computationally very demanding for chains of realistic lengths.

Deutsch and Kurosky (DK) attempted to estimate the partition function in high-temperature approximation taking into account the first cumulant only by presenting the partition function Z in the simplest form:

$$F_s = -T \ln Z = - \sum_{1 \leq i < j \leq N} (U(\sigma_i, \sigma_j)) < \Delta(r_i, r_j) > \quad (12)$$

where the $< >$ denote unbiased averaging over all conformations.

It is quite clear that for compact chains the approach of DK is basically equivalent to the earlier approach in [9] that assumed sequence independence of the partition function. Indeed in globular polymers the $< \Delta_{ij} >$ (which has the physical meaning of the probability of a contact between monomers i and j in the full ensemble of conformations) does not depend on i and j except when these monomers are close to each other along the chain [35,57]. It is clear that setting $< \Delta_{ij} > = \text{const}$ in eq.(12) results in sequence independence

of the partition function. In apparent contradiction with the above arguments DK reported a considerable improvement (for the 2-letter HP model) over the results of the previous approach [9].

It is possible that the improvement over the simplest design reported in [51] is due to the special property of the cubic lattice that excludes the contacts for which $j - j$ is even. In other words *on a cubic lattice* $\langle \Delta_{ij} \rangle \approx \text{const}$ when $i - j$ is odd and is 0 otherwise. The design in [51] took advantage of this property of the cubic lattice providing proper distribution of H and P monomers over even or odd sites.

It is also worth mentioning that both Seno et al and DK used the HP model to test the results of their design procedures. In both cases the methodologies are not limited technically to the HP model. As was explained before, the HP model is problematic to study design and folding. For the two-letter model on the square lattice (as well as on the cubic lattice with average attraction between monomers) $m_{eff} \approx \gamma$, i.e. it is on the verge of failure. That makes the design results for the HP model unstable and heavily dependent on the details of a model such as lattice type, chain length, "even-odd" contacts, details of the composition etc. It is quite possible that some improvements of the design methods over the simplest one suggested in [9] actually solve the problems specific to the gapless HP model. Those problems may not exist in more realistic multiple-letter models, where any reasonably compact structure is designable even within the simplest algorithm of [9].

To this end it would be desirable to apply interesting design methods proposed by DK and Seno *et al* to 20 aminoacid types model and compare folding rates and stability of sequences designed using various procedures.

Morrissey and Shakhnovich (MS) [52] proposed a new design procedure which seeks sequences having high probability to be in their native state at a given temperature T , $P(T)$. This procedure also employs MC in sequence space; however the partition function of the chain Z entering the expression for $P(T)$ in eq.(9) is estimated using the cumulant expansion approximation. This eliminates the need to run simulations in conformational space after each mutation to estimate the partition function [53] and thus dramatically

increases the computational efficiency.

This design procedure was carried out for 20-letter model proteins of various sizes (36-mers and 64-mers) on a cubic lattice and turned out to be quite efficient yielding sequences that are stable at a selected temperature. Two interesting and unexpected results emerged from this study: First, the *folding transition temperature* for designed sequences turned out to be highly correlated with the input temperature at which designed sequences were stable in their native conformations.

Second, the temperature at which folding *rate* was the fastest, appeared to be very close to the stability temperature T which was input in the algorithm. This reflects an important feature of proteins that optimum of their folding kinetics is achieved at the conditions when their native state is not extremely stable - a finding fully consistent with the well-known marginal stability of natural proteins. The reason for such relation between thermodynamics and kinetics is partly given in a simple theory of folding kinetics presented in [32].

The observed correlation between folding rate and folding temperature generates an interesting prediction that proteins from thermophilic organisms should fold very slow at normal temperature (around $300K$) at which folding of mesophilic proteins is fast. This prediction is partly supported by the observation that some thermophilic proteins (e.g. ribonucleotide Reductase from *ThermusX - 1* [58]) are most active at high temperature (about $90C$) and they retain only marginal activity at room temperature. The implicit assumption made here is that enzymatic activity correlates with foldability. The validity of this assumption requires further study.

Interestingly, different features of folding sequences were emphasized in the MS procedure at different input folding temperatures. Sequences that were designed to be stable at high T featured low energy in the native state and higher dispersion of interaction energies D . In contrast, sequences that were designed to fold at lower temperature had lower D and higher E_N (see Fig.11 of [52]). This result shows that an optimal design strategy may be different for design of thermostable and mesophile sequences. A possible reason for that was discussed in [52].

V. DESIGNING LONGER SEQUENCES THAT FOLD COOPERATIVELY.

The theoretical approaches to protein design were based on the results of mean-field heteropolymer theory which did not take into account inhomogeneity in the distribution of interacting aminoacids over the protein structure. This approximation neglects the fact that some parts of the protein, e.g. interior may have been stabilized to a greater extent than other parts, e.g. exterior. Lattice simulation showed that this factor may be important for longer proteins giving rise to a "multidomain" behavior where core folds at higher temperature than the surrounding loops, leading to lower folding cooperativity [59–61]. It was shown [59,62,61] that existence of domains is correlated with δ , the dispersion of *native* contact energies. Sequences having higher δ tend to fold less cooperatively (core first, then loops) while sequences with lower δ fold as a one cooperative unit. An improved design procedure which optimizes both Z -score and δ was proposed in [62]. This approach makes it possible to design sequences having desired folding cooperativity.

VI. EVOLUTION-LIKE DESIGN OF FAST-FOLDING SEQUENCES

Thermal stability is not the only feature of protein sequences that could be optimized. Another important characteristic is folding rate. It is of great interest to compare the sequences optimized for stability with the ones optimized for folding rates because it may shed some light on the features of proteins that were optimized in natural evolution of their sequences. The evolution-like selection of fast-folding sequences was suggested in [63] and further developed in [64]. The idea of the method is conceptually simple and similar to the design that optimizes the stability. Mutations are attempted and only those are accepted that make folding faster (details are in [63,64]). The algorithm has proven successful yielding many fast-folding sequences. Analysis of the "database" of emerged sequences showed that they are indeed more thermodynamically stable in their native conformations, than random sequences. Interestingly, the Z -scores of evolved fast folding sequences were markedly lower

than for random sequences but markedly higher than for sequences that were designed to optimize their Z -score (we remind the reader that Z scores are always negative, i.e. "lower" means "better", as far as stability is concerned). Despite of higher Z -score, sequences generated by evolution-like selection procedure folded much faster than sequences designed for higher stability (an order of magnitude at the respective temperatures of fastest folding). This points out clearly to the usefulness and limitation of the Z -score as predictor of the folding rate (as well as any other global thermodynamic criterion).

A more detailed analysis of the features of evolved fast-folding sequences showed that their stabilizing interactions were distributed unevenly: acceleration of folding was accompanied by stabilization of specific fragment of the structure (the "folding nucleus" [65–68,3,4]), while the remaining part of the structure was much less stabilized. In other words, in the evolution-like selection of fast-folding sequences the first few mutations lead to the decrease of Z -score accompanied by some acceleration of folding. Further acceleration was achieved after a few subsequent mutations that strengthened specific set of contacts, the folding nucleus. In the steady state of evolution-like selection where folding rate did not change much with mutations the aminoacids at the nucleus positions were remarkably conserved in contrast to other positions where mutations were frequent.

A similar approach was taken by Nadler and coworkers in their interesting study of 2-dimensional protein model [69]. These authors pointed out that in their model the energy optimization does not always give the desired results and additional optimization of folding rate may be required to find folding sequences. This conclusion is consistent with the theoretical views presented in this review (see e.g. Lesson 4): Two-dimensional models behave very differently and the results obtained with these models cannot be directly compared with the results from three-dimensional models. To understand better the differences between two-dimensional models and three-dimensional ones it is of clear interest to study the features of sequences selected for fast folding in [69]

VII. LESSONS FOR FOLDING

The best and most objective criterion of success in protein design is folding of designed sequences, *in vitro*, or *in vivo* or *in silico*. Clearly, certain features of the folding phenomenology depends crucially on how the sequences were designed/selected. This fact calls for great caution in comparing folding in different models where sequences were designed (selected) using different methods. In particular sequences that have large energy gap $E_N - E_c$ fold cooperatively ("first order like"). In contrast, weakly designed or random heteropolymers that do not have such a large gap, have non-cooperative folding transition. [14,1,19]). Another examples show that such features as on [59] and off-pathway [70,54] intermediates may be designed "in" or "out" by proper sequence selection.

E.g. the folding dynamics for two sequences designed to fold into the same 36-mer conformation but using different design strategies were compared in [54] The first sequence, *Seq1* was designed by optimizing the Z -score (at a variable aminoacid composition) while the second one, *Seq2* was generated using the original approach [9] that minimizes the native state energy at constant aminoacid composition. It was shown that the sequence *Seq1* that was obtained by optimizing the Z -score folded fast, more cooperatively and was more stable in the native state than *Seq2*. While the transition for *Seq1* followed the two-state scenario both in thermodynamics and kinetics, an equilibrium intermediate and structurally similar to it trapped kinetic intermediate were found for *Seq2*.

Since both thermodynamics and kinetics are derived from the properties of the energy landscape there is an established relation between them (see e.g. [71]). To this end care should be taken in comparing the results of folding simulations for different models in which sequences were designed differently. Such comparison is possible only if equilibrium behavior of two models are similar. E.g. recent studies [72] showed that folding transition in some off-lattice models is non-cooperative in contrast to lattice models and experiment [18,73,74]. This fact rules out the nucleation mechanism for the model of Ref. [75]. Correspondingly it may be not very insightful to compare the cooperative kinetics of real proteins and lattice

model proteins with the non-cooperative kinetics in the off-lattice model studied in [75,72,76].

The theoretical developments in protein design stimulated interesting experimental studies including design with reduced, or simplified alphabets to address the issue of a "minimalistic" protein sequence, i.e. what is the minimal number of amino acid types that make it possible to design stable folding sequences. Hecht and coworkers [77] designed and synthesized sequences based on the "two-aminoacid type" assumption that distribution of hydrophobic aminoacids is most crucial determinant of the structure. While thus designed proteins were compact and belonged to the expected (helical) secondary structure class, their folding into unique structure and cooperativity has not been fully established. In a recent elegant study by Baker and coworkers [7] the phage display technique was employed to seek "minimalistic" sequences that fold into the structure of a small protein, SH3, as judged by its activity. The authors of [7] come to the conclusion that 6 aminoacids alphabet is generally sufficient for protein design, with an important exception of a few sites where simplification was not possible. One possibility is that these sites are related to function, another possibility that they participate in the unique folding nucleus. Future studies will clarify this important issue.

VIII. CONCLUDING REMARKS

One of the main points of this review is that better understanding of protein folding (at least in the realm of simple models) is of crucial importance to the success of protein design.

The results of statistical-mechanical analysis (see eqs(3,5,8) and Lesson 1) show that for an appropriate model (for which $m_{eff} > \gamma$) exponentially (in chain length N) large number of sequences, can fold cooperatively into a given structure. This is consistent with the observation that many non-homologous protein sequences can fold into similar conformations [78], the fact that makes the "bioinformatics" approach to prediction of protein conformation so difficult. From the design perspective, the chance that designed sequence is identical or even homologous to the native sequence is minimal. Therefore the success of design cannot

be measured by relatedness of "predicted" and native primary structure [46]. However, when aminoacids are categorized into small number of classes, the simplest division being into hydrophobic and polar the correlation between "predicted" and real sequences is beyond the noise level [41]. However, as was noted earlier the models that have only two kinds of aminoacids essentially fail to fold (unless the ensemble of conformations is very restricted).

It is almost tautological to say that design represents a search in sequence space to optimize folding and stability. The straightforward approaches this problem that directly (from simulations) evaluate the impact of each mutation on folding thermodynamics [53] or kinetics [63,69], are computationally very intensive and at that point are hardly feasible for models other than simplest lattice model. This calls for a powerful folding criterion that is easy to evaluate without running simulations in conformational space after each mutation. Such criterion should be a good predictor of folding ability that can be used as a "scoring function" to be optimized in sequence space. Here the theory of folding provides a crucial contribution to design pointing out to such criteria as energy gap and related to it Z -score as well as δ , the dispersion of energies of native contacts and in some cases the stability of the nucleus. Importantly those criteria are correlated to stability and folding rate (in a certain range of temperatures, see [52,32]) and therefore they proved very useful for design. A useful folding criterion should be simple and easy to evaluate without intensive searches in conformational space. E.g. recently, the so-called σ -criterion was proposed to distinguish between fast- and slow folding sequences [11]. While in essence this criterion is related to the Z -score, or gap criterion ([4], A.Dinner,M.Karplus and ES, to be published) its value is not known without the folding simulations. That makes the utility of the σ -criterion for protein design problematic.

Obviously the folding criteria that are currently used for design have their limitations. In particular there is evidence that fast folding could have been an important factor in evolutionary selection of proteins [79,64]. This may call for a criterion that takes the folding kinetics into account more consistently (a step in this direction was outlined in [80]). It is likely that search for better simple folding criteria will remain an important area of research

at the interface between protein folding and design.

Another crucial bottleneck in protein design is the lack of knowledge of a potential function that faithfully reproduces protein energetics (i.e. for which the native structure for the native sequence is global energy minimum with energy gap). This direction of research has been extremely active (see e.g. in [81,82,8,27]) and is likely to be very active in future. The major issue here is to find a model that is still feasible to simulate but which has enough detail to make it possible to derive "good" folding potentials. It was shown in [27,83] that simple pairwise contact potential approximation is too crude to describe real proteins. There is no set of parameters U that provides energy gap that is sufficient for successful folding simulations of real proteins, in the two-body contact approximations of the energetics. It is almost certain that future studies will seek better potentials for more refined models (see e.g. [81,84]) that can be used for reliable design approaches.

A crucial direction of the further study is to bring the progress in theoretical protein design closer to experiment. An important issue that needs to be addressed in applying theoretical models to the design of real proteins is whether the details of side-chain packing are crucial determinants of a protein structure. While some original proposals gave affirmative answer to this question [85,86] more recent experimental studies indicated that chain flexibility needs to be taken into account so that many side-chains substitutions can be accommodated by slightly varying the backbone conformations [87,88]. Interesting methods to account for side-chain stereochemistry in sequence selections have been developed [89,6,90] that use dead-end elimination theorem or Monte-Carlo design that takes into account side chains degrees of freedom [91].

An important signature of the maturity of a field is the degree of interaction between theory and experiments. By that criterion protein design enters its maturity stage and we are entitled to witness stunning progress in the near future.

ACKNOWLEDGMENTS

I am grateful to Alex Gutin, Victor Abkevich, Leo Mirny and Mike Morrissey for their numerous contributions and insights. This work was supported by NIH grant GM52126

FIGURE LEGENDS

1 A schematic presentation of the protein design problem (taken from [52]): given the target 3D structure and the selected temperature find a sequence that folds at this temperature into the given conformation and is stable in this conformation.

2 Degeneracy of the protein code. The solid line is the analytical formula (3). The average E_{av} and dispersion D was calculated as explained in the text using the Myazawa-Jernigan set of parameters (table VI [22]). Simulations using other parameter set [23] provided identical results.

Data points correspond to the direct calculation of sequence entropy from MC simulations in a range of selective temperatures (keeping the aminoacid composition same as in the native sequence). Average energy of sequences in the target structure $E(T)$ was evaluated from simulation runs Then eq.(11) was applied to obtain sequence space entropy. (Here we show entropy and energy , normalized per aminoacid residue $s_{seq} = S_{seq}/N$, $e_N = E_N/N$). Different symbols correspond to different proteins (in pdb access code): filled diamonds - 4mbn, open squares - 2cab, filled squares - 1pcy, open diamonds - 2pal. Horizontal insert is given for illustrative purpose to show schematically the generic representation of density of states in conformational space, as predicted by the heteropolymer theory [14,19]. The range of energies at which density of non-native decoys is high is shown in black, a few low energy conformations (shown as discrete lines in the insert) that lie below the boundary of continuous spectrum E_c represent lowest energy decoys.

a) "Designable model" where $m_{eff} > \gamma$. Many sequences ($\sim exp(1.9N)$ in the present example) exist that have low energy E_N in the target conformation with pronounced stability gap $\Delta = E_N - E_c$. Such sequences are expected to fold fast into the native conformation

b) Non-designable model $m_{eff} < \gamma$: no sequences that fold uniquely to the ground state can be found. The model runs out of sequences at energies which are not low enough to ensure large gap between the native structure and misfolded decoys. The data points represent MC design simulations entropy for two "HP" models of proteins: 1mbn (upper

curve) and 1pcy (lower curve). Aminoacids were categorized into "H" and "P" classes as explained in [41]. The more pronounced difference between proteins is due to the difference in their average hydrophobicities. i.e fraction of hydrophobic residues in their sequences.

3 The density of states (energy spectra) in the ensemble of fully compact conformations of the 27-mer model for a random and best designed sequence. Each bar corresponds to entropy per residue - the logarithm of the number of all conformations having given energy divided by the number of residues (27 in this case). The density of states plots are derived from exhaustive enumeration of all 103346 compact conformations of the 27-mer [29]. For simplicity only two types of monomers are used ("black" and "white") with nearest neighbor "color specific" interactions: $E_{BB} = E_{WW} = -3$; $E_{BW} = -1$ [9,25]. While this interaction matrix may be not quite realistic for real proteins, it is useful for clarifying basic concepts presented in this review. Obviously the lowest energy conformation is the one that maximizes the number of favorable "same color" (SC) contacts. Left insert shows the target structure and the sequence that has minimal possible energy $E_{lowest} = -84$ (all 28 contacts are SC) in that structure. This structure represents a unique ground state for the designed sequence: The black bar for the designed sequence corresponding to the energy $E_N = -84$ is slightly exaggerated to make it visible. The right insert shows the same structure with a quasirandom sequence fit into it.

REFERENCES

- [1] M.Karplus & E.Shakhnovich. *Protein Folding*, chapter 4, pages 127–196. W.H.Freeman and Company, New York, (1992).
- [2] J.Bryngelson, J.N.Onuchic, N.D.Socci, & P.Wolynes. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis *Proteins: Struct. Funct. and Genetics* **21**, 167–195.
- [3] A.Fersht. (1997). Nucleation mechanism of protein folding *Curr. Opin. Struct. Biol.* **7**, 10–14.
- [4] E.I.Shakhnovich. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29–40.
- [5] T.Quinn, N.Tweedy, R.Williams, J.Richardson, & D.Richardson. (1994). Betadoublet: de novo design, synthesis and characterization of a β -sandwich protein *Proc Natl. Acad. Sci. USA* **91**, 8747–8751.
- [6] B.Dahiyat & S.Mayo. (1997). De nove design: fully automated sequence selection *Science* **278**, 82–87.
- [7] N.S. Riddle, J.V.Santiago, S.T.Bray, N.Doshi, V.Grantchanova, Q.Yi, & D.Baker. (1997). Functional rapidly folding proteins from simplified aminoacid sequences *Nature Structural Biology* **4**, 805–809.
- [8] R.Goldstein, Z.A. Luthey-Schulten, & P.Wolynes. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
- [9] E.Shakhnovich & A.Gutin. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc.Natl. Acad. Sci. USA* **90**, 7195–7199.
- [10] A.Sali, E.I.Shakhnovich, & M.Karplus. (1994). Kinetics of protein folding. a lattice model study for the requirements for folding to the native state. *Journ. Mol. Biol.* **235**,

- 1614–1636.
- [11] D.Klimov & D.Thirumalai. (1996). A criterion which determines foldability of proteins *Phys.Rev.Lett* **76**, 4070–4073.
 - [12] C.Anfinsen. (1973). Principles that govern the folding of protein chains *Science* **181**, 223–230.
 - [13] Y.Ueda H.Taketomi & N.Go. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. *Intl.Journal Peptide Prot.Res.* **7**, 445–449.
 - [14] E.I.Shakhnovich & A.M.Gutin. (1989). Formation of unique structure in polypeptide chains. theoretical investigation with the aid of replica approach. *Biophysical Chemistry* **34**, 187–199.
 - [15] C.Sfatos, A.M.Gutin, & E.I.Shakhnovich. (1993). Phase diagram of random copolymers *Phys. Rev.E* **48**, 465.
 - [16] S.Ramanathan & E.Shakhnovich. (1994). Statistical mechanics of proteins with “evolutionary selected” sequences *Phys.Rev.E* **50**, 1303–1312.
 - [17] V.Pande, A.Yu. Grosberg, & T.Tanaka. (1995). Freezing transition of random heteropolymers consisting of arbitrary sets of monomers *Phys.Rev.E* **51**, 3381–3393.
 - [18] E.I.Shakhnovich. (1994). Proteins with selected sequences fold to their unique native conformation *Phys.Rev.Lett.* **72**, 3907–3910.
 - [19] V.Pande, A.Grosberg, & T.Tanaka. (1997). Statistical mechanics of simple models of protein folding and design *Biophysical Journal* **73**, 3192–3210.
 - [20] V.S.Pande, A.Yu.Grosberg, & T.Tanaka. (1994). Folding thermodynamics and kinetics of imprinted renaturable heteropolymers *Journal of Chemical Physics* **101**, 8246–8257.
 - [21] J.D.Bryngelson & P.G.Wolynes. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc.Natl.Acad.Sci.USA* **84**, 7524–7528.

- [22] S.Myazawa & R.Jernigan. (1985). Estimation of effective interresidue contact enrgies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552.
- [23] A.Kolinski, A.Godzik, & J.Skolnick. (1993). The general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J.Chem.Phys.* **98**, 7420–7433.
- [24] E.I.Shakhnovich, G.M.Farztdinov, A.M.Gutin, & M.Karplus. (1991). Protein folding bottlemnecks: A lattice monte-carlo simulation *Phys.Rev.Lett.* **67**, 1665–1667.
- [25] N.Socci & J.Onuchic. (1994). Folding kinetics of protein like heteropolymers *J.Chem.Phys.* **101**, 1519–1528.
- [26] M.Sippl. (1990). Calculation of conformational ensemble from potential of mean force. an approach to knowledge-based prediction of local structures in globular proteins *J.Mol.Biol.* **213**, 859–883.
- [27] L.Mirny & E.Shakhnovich. (1996). How to determine protein folding potential? a new approach to the old problem. *J.Mol.Biol* **264**, 1164–1169.
- [28] A.V.Finkelstein, A.Gutin, & A.Badretdinov. (1993). Why are some protein structures so common? *FEBS Lett.* **325**, 23–28.
- [29] E.I.Shakhnovich & A.M.Gutin. (1990). Implications if thermodynamics of protein folding for evolution of primary sequences. *Nature* **346**, 773–775.
- [30] A.M.Gutin & E.I.Shakhnovich. (1993). Ground state of random copolymers and the discrete random energy model *J.Chem.Phys* **98**, 8174–8177.
- [31] M.Mezard, G.Parisi, & M.Virasoro. *Spin Glass Theory and Beyond*. World Sci., Singapore, (1988).
- [32] A.Gutin, A.Sali, V.Abkevich, M.Karplus, & E.Shakhnovich. (1998). Temperature de-

- pendence of folding in a simple proteinlike model: Search for glass transition *J. Chem. Phys*, *in press* pages xxx–xxx.
- [33] K.Yue, K.Fiedig, P.Thomas, H.S.Chan, E.I.Shakhnovich, & K.A.Dill. (1995). A test of lattice protein folding algorithms *Proc. Natl. Acad. Sci. USA* **92**, 325–329.
 - [34] E.M.O’Toole & A.Z.Panagiotopoulos. (1993). Effect of sequence and intermolecular interactions on the number and nature of low-energy states of simple model proteins. *J. Chem. Phys.* **98**, 3185–3190.
 - [35] A.Yu. Grosberg & A.R.Khohlov. *Statistical Mechanics of Macromolecules*. AIP Press, N.Y.,N.Y, (1994).
 - [36] V.Abkevich, A. Gutin, & E.Shakhnovich. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding *Journ.Mol.Biol.* **252**, 460–471.
 - [37] E.I.Shakhnovich & A.M.Gutin. (1989). Frozen states of disordered globular heteropolymers. *J.Phys* **A22**, 1647.
 - [38] V.Pande, A.Grosberg, C.Joerg, & T.Tanaka. (1996). Is heteropolymer freezing well described by the random energy model? *Phys Rev Lett* **76**, 3987–3990.
 - [39] S. Govindarajan & R.Goldstein. (1995). Searching for foldable protein structures using optimized energy functions *Biopolymers* **36**, 43–51.
 - [40] A.Yu.Grosberg, S.K.Nechaev, & E.I.Shakhnovich. (1988). The role of topological constraints in the kinetics of collapse of macromolecules *J.Physique (France)* **49**, 2095–2100.
 - [41] E.Shakhnovich & A.Gutin. (1993). A novel approach to design of stable proteins. *Protein Engineering* **6**, 793–800.
 - [42] V.Pande, A.Yu. Grosberg, & T.Tanaka. (1994). Thermodynamic procedure to synthesize heteropolymers that can renature to recognize a given target molecule *Proc. Natl. Acad. Sci. USA* **91**, 12976–12979.

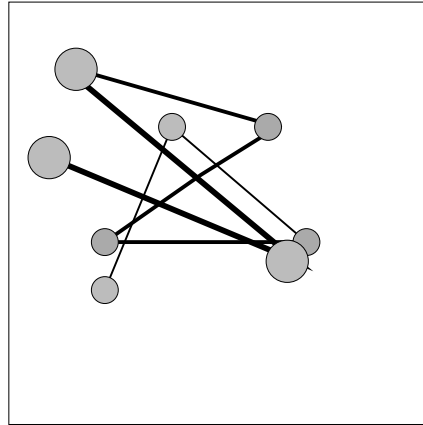
- [43] A. Gutin, V. Abkevich, & E. Shakhnovich. (1995). Is burst hydrophobic collapse necessary for rapid folding? *Biochemistry* **34**, 3066–3076.
- [44] M. Chung, A. Neuwald, & W. J. Wilbur. (1998). A free energy analysis by unfolding applied to 125-mers on a cubic lattice *Folding & Design* **3**, 51–65.
- [45] D. Jones. (1995). Theoretical approaches to designing novel sequences to fit a given fold *Curr Opin Biotechnology* **6**, 452–459.
- [46] P. Koehl & M. Delarue. (1996). Mean-field minimisation methods for biological macromolecules *Curr Opin Struct Biol* **6**, 222–226.
- [47] J. Saven & P. Wolynes. (1997). Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules *J. Phys. Chem.* **101**, 8375–8389.
- [48] A. V. Finkelstein, A. Gutin, & A. Badretdinov. (1995). Why are the same protein folds used to perform different functions? *Proteins: Struct. Function genetics* **23**, 142–149.
- [49] H. Li, N. Winfree, & C. Tang. (1996). Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669.
- [50] E. I. Shakhnovich & A. M. Gutin. (1990). Exhaustive enumeration of all conformations of compact heteropolymers with quenched disordered sequence of links *J. Chem. Phys* **93**, 5967–5971.
- [51] J. M. Deutsch & T. Kurosky. (1996). New algorithm for protein design *Phys. Rev. Lett.* **76**, 323–326.
- [52] M. Morrissey & E. Shakhnovich. (1996). Design of proteins with selected thermal properties *Folding & Design* **1**, 391–406.
- [53] F. Seno, M. Vendruscolo, A. Maritan, & J. Banavar. (1996). Optimal protein design procedure *Phys. Rev. Lett.* **77**, 1901–1904.
- [54] L. Mirny, V. Abkevich, & E. Shakhnovich. (1996). Universality and diversity of the protein

- folding scenarios: A comprehensive analysis with the aid of lattice model. *Folding & Design* **1**, 103–116.
- [55] J.U.Bowie, R.Luthy, & D.Eisenberg. (1991). A method to identify protein sequences that fold into a known three-dimensional structure *Science* **253**, 164–169.
- [56] A.Gutin, V.Abkevich, & E.Shakhnovich. (1996). Chain length scaling of protein folding time *Phys Rev Lett* **77**, 5433.
- [57] E.Shakhnovich. *Statistical Mechanics, Protein Structure and Protein-Ligand Interactions*. Plenum, New York, (1994).
- [58] G.N.Sando & P.C.Hogenkamp. (1973). Ribonucleotide reductase from *thermus x1*, a thermophilic organism *Biochemistry* **12**, 3316–3322.
- [59] V.Abkevich, A. Gutin, & E.Shakhnovich. (1995). Domains in folding of model proteins *Protein Science* **4**, 1167–1177.
- [60] A. Gutin, V.Abkevich, & E.Shakhnovich. (1998). Cooperativity of protein folding and the random-field ising model *Phys Rev E* pages xxx–xxx.
- [61] A.Panchenko, Z. Luthey-Schulten, & P.Wolynes. (1995). Foldons, protein structural modules and exons *Proc. Natl. Acad. Sci. USA* **93**, 2008–2013.
- [62] V.Abkevich, A. Gutin, & E.Shakhnovich. (1996). Improved design of stable and fast-folding proteins. *Folding & Design* **1**, 221–232.
- [63] A. Gutin, V.Abkevich, & E.Shakhnovich. (1995). Evolution-like selection of fast-folding model proteins *Proc Natl. Acad. Sci. USA* **92**, 1282–1286.
- [64] L.Mirny, V.Abkevich, & E.Shakhnovich. (1998). How evolution makes proteins fold quickly *Proc Natl. Acad. Sci. USA* pages xxx–xxx.
- [65] V.Abkevich, A. Gutin, & E.Shakhnovich. (1994). Specific nucleus as the transition state for protein folding: Evidence from the lattice model *Biochemistry* **33**, 10026–10036.

- [66] L. Itzhaki, D.Otzen, & A.Fersht. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding *J.Mol.Biol.* **254**, 260–288.
- [67] A.R.Fersht. (1995). Optimization of rates of protein folding: The nucleation-condensation mechanism and its implications *Proc. Natl. Acad. Sci. USA* **92**, 10869–10873.
- [68] E.Shakhnovich, V. Abkevich, & O.Ptitsyn. (1996). Conserved residues and the mechanism of protein folding *Nature* **379**, 96–98.
- [69] M.Ebeling & W.Nadler. (1995). On constructing folding heteropolymers *Proc.Natl. Acad. Sci. USA* **92**, 8798–8802.
- [70] V.Abkevich, A. Gutin, & E.Shakhnovich. (1994). Free energy landscape for protein folding kinetics. intermediates, traps and multiple pathways in theory and lattice model simulations. *J.Chem.Phys* **101**, 6052–6062.
- [71] E.M.Lifshits & L.P.Pitaevskii. *Physical Kinetics*. Pergamon, Oxford; New York, (1981).
- [72] Z.Guo & C.Brooks. (1997). Thermodynamics of protein folding: A statistical-mechanical study of a small all β -protein *Biopolymers* **42**, 745–757.
- [73] N.Socci & J.Onuchic. (1995). Kinetics and thermodynamic analysis of proteinlike heteropolymer: Monte carlo histogram technique *J.Chem.Phys.* **103**, 4732–4744.
- [74] P.L.Privalov. (1996). Intermediate states in protein folding *Journ. Mol. Biol.* **258**, 707–725.
- [75] Z.Guo & D.Thirumalai. (1995). Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labelling experiments. *Biopolymers* **35**, 137–139.
- [76] Z.Guo & D.Thirumalai. (1997). The nucleation collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Folding & Design* **2**, 377–391.

- [77] M.Kamtekar, M.Schiffer, H.Xiong, J.Babik, & M.Hecht. (1993). Protein design by binary patterning of polar and nonpolar aminoacids *Science* **262**, 1680–1685.
- [78] L.Holm & C.Sander. (1993). *J.Mol.Biol* **233**, 123–138.
- [79] A.Ladurner & L. Itzhaki A.Fersht. (1997). Strain in the folding nucleus of chymotrypsin inhibitor 2 *Folding & Design* **2**, 363–366.
- [80] V.S.Pande, A.Yu.Grosberg, D.Rokhsar, & T.Tanaka. (1998). Pathways for protein folding: is a “new view” needed *Curr Opin Struct Biology, in press* **8**, xx–xx.
- [81] R.Jernigan & I.Bahar. (1996). Structure-derived potentials and folding simulations *Curr Opin. Struct. Biol.* **6**, 195–209.
- [82] D.Jones & J.Thornton. (1996). Potential energy functions for threading *Curr Opin. Struct. Biol.* **6**, 210–216.
- [83] M.Vendruscolo & E.Domany. (1998). Elusive unfoldability: Learning a contact potential to fold crambin *J.Mol.Biol*, *submitted*.
- [84] R.S. DeWitte & E.I.Shakhnovich. (1996). Smog: de novo design method based on simple, fast and accurate free energy estimates. 1. methodology and supporting evidence *J.Amer.Chem.Soc* **118**, 11733–11744.
- [85] J.Ponder & F.Richards. (1994). Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes *J.Mol.Biol* **193**, 5803–5807.
- [86] W.Lim & R.Sauer. (1991). The role of internal packing interactions in determining the structure and stability of a protein *J.Mol.Biol.* **219**, 359–376.
- [87] W.Lim, A.Hadel, R.Sauer, & F.Richards. (1994). The crystal structure of a mutant protein with altered but improved hydrophobic core *Proc Natl. Acad. Sci. USA* **91**, 423–427.

- [88] E.Baldwin, O.Hajiseyedjavadi, W.Baas, & B.Mathews. (1993). The role of backbone flexibility in the accomodation of variants that repack the core of t4 lysozyme *Science* **262**, 1715–1718.
- [89] B.Dahiyat & S.Mayo. (1995). Probing the role of packing specificity in protein design *Proc. Natl. Acad. Sci. USA* **94**, 10172–10177.
- [90] M. De Maeyer, J.Desmet, & I.Lasters. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination *Folding & Design* **2**, 53–66.
- [91] H.Helinga & F.Richards. (1994). Optimal sequence selection in proteins of known structure by simulated evolution *Proc. Natl. Acad. Sci. USA* **91**, 5803–5807.



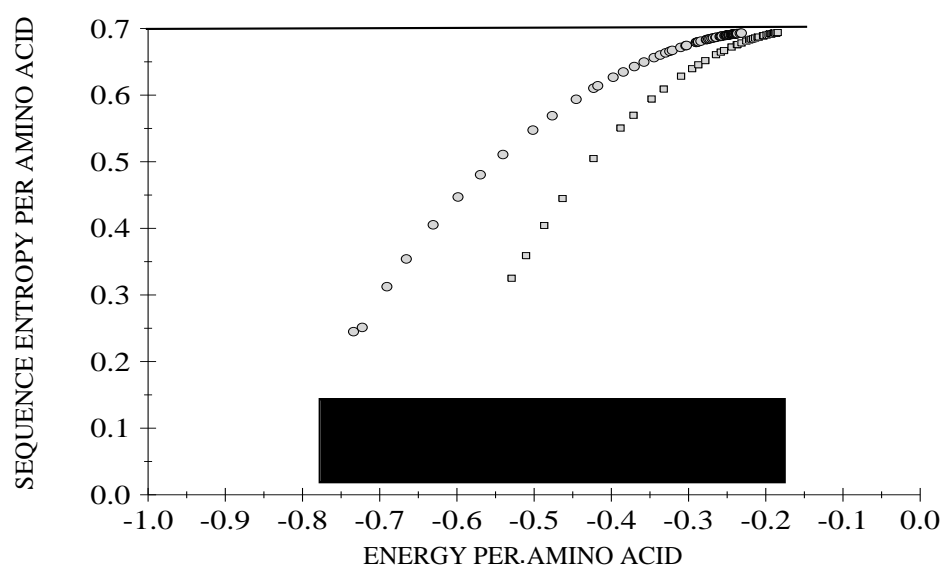
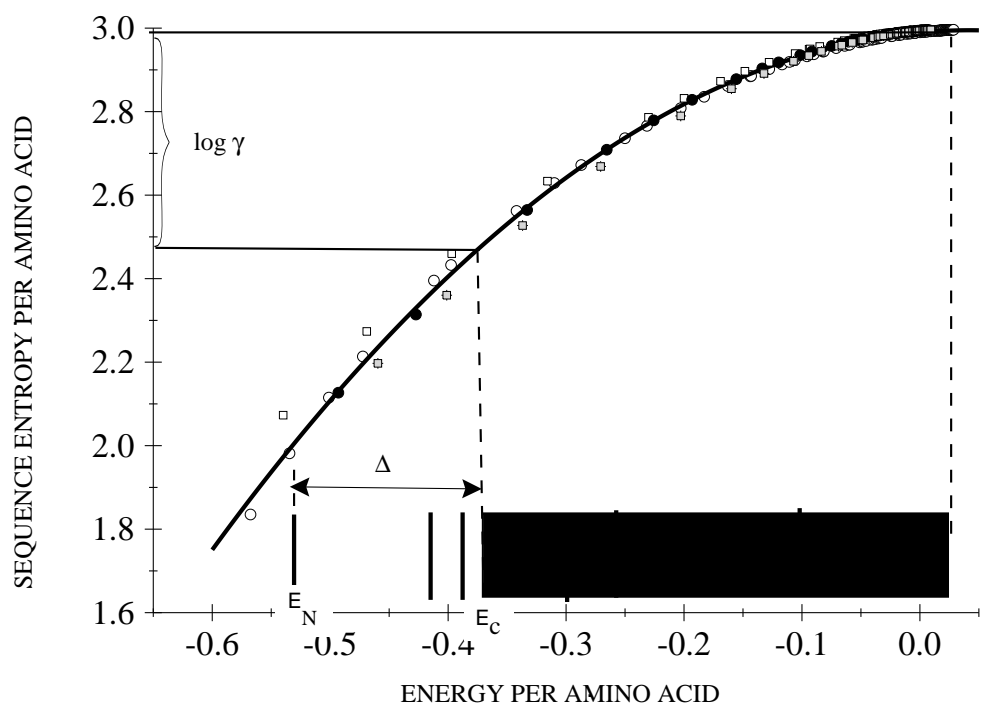
3-D Target Structure

T

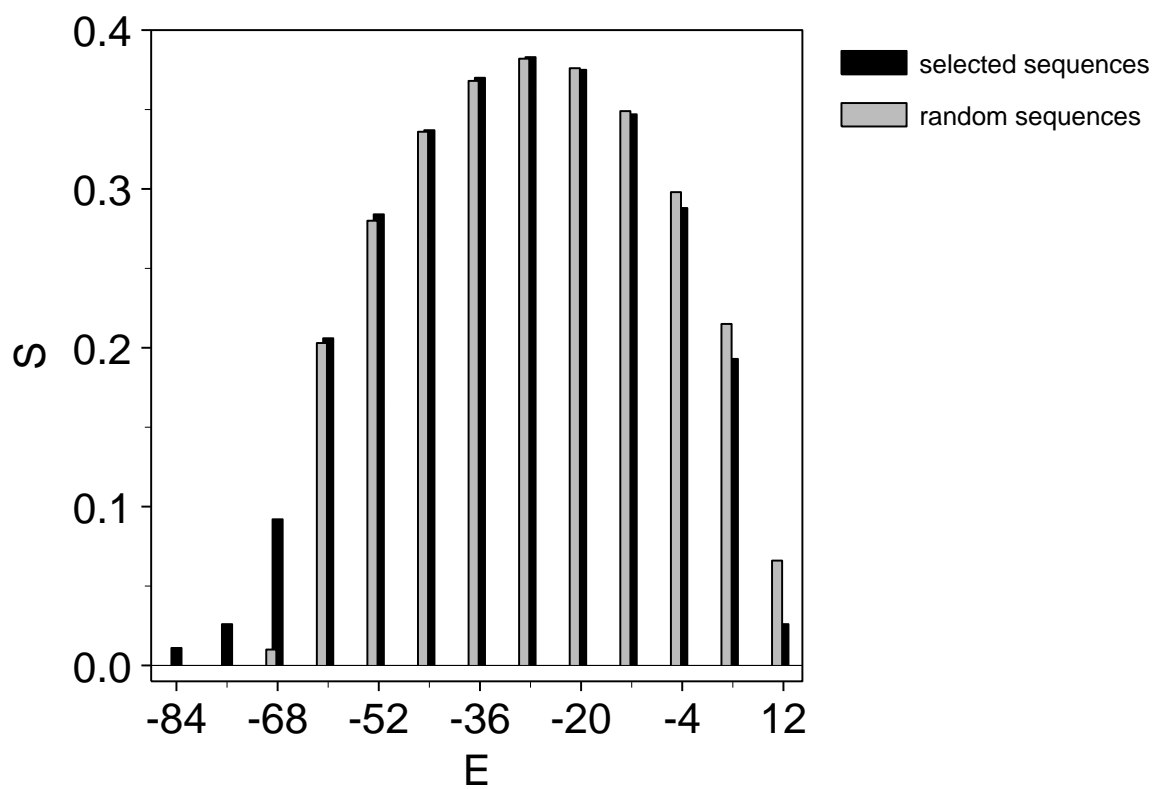
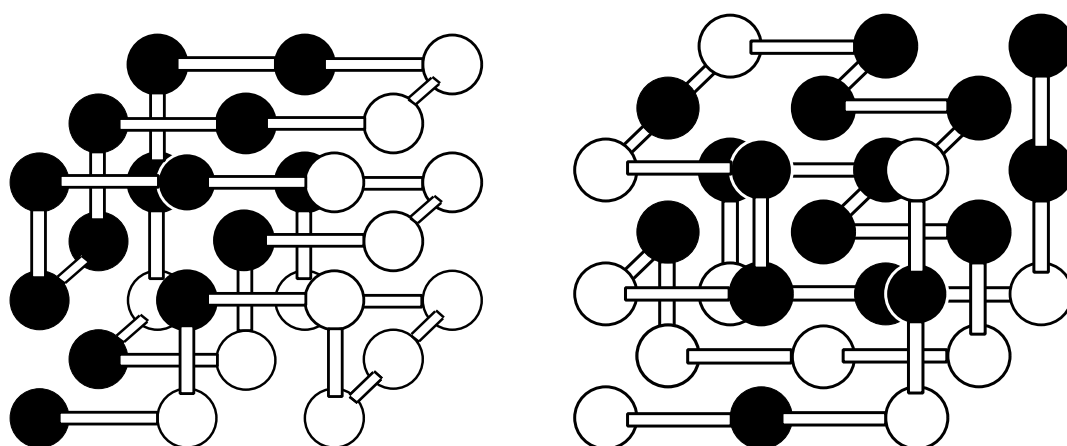
Design Temperature

L-T-G-C-I-P-Q-W

Sequence which folds to target structure



"Protein design..". Fig.2



"Protein design..". Fig.3